Indian J. Phys. **80** (4), 347-355 (2006)

Direct phasing of one-wavelength anomalous scattering data : a high throughput tool in structural genomics

S Selvanayagam,[1] D Velmurugan[1,*] and T Yamane[2]

[1]Department of Crystallography and Biophysics, University of Madras, Guindy Campus, Chennai – 600 025, India

[2]Department of Biotechnology and Biomaterial Science, Graduate School of Engineering, Nagoya University, Furo-Cho, Chikusa-Ku, Nagoya 464-8603, Japan

E-mail d_velu@yahoo.com

Received 27 October 2005, accepted 25 February 2006

Abstract The available macromolecular sequence information exceeds far in number the available three-dimensional structures. High throughput techniques are hence necessary to unravel the three-dimensional structures of selected macromolecular sequences in the area of Structural Genomics in a short time. The structure solution program SHELXD is useful for locating the anomalous scatterers from SIR, SAS, SIRAS or MAD data. SHELXE uses the native phases and the weights from SHELXD. OASIS is a computer program for breaking phase ambiguity in one wavelength anomalous scattering data. The phases obtained from SHELXE and OASIS are of superb quality to allow automated model building to be carried out in ARP/wARP. Attempts are here made in extending the applications to the high throughput structure elucidation of thermolysin of approximately 34 kDa molecular weight using 1.7 Å single wavelength anomalous scattering (SAS) data and 2 Å truncated data and also of glucose isomerase of approximately 44 kDa molecular weight using 1.45 Å SAS data.

Keywords SAS, glucose isomerase, thermolysin

PACS Nos. 61.10.Nz, 87.14.Ee

1. Introduction

The High Throughput Crystallography Consortium was developed to refine and extend the powerful software tools that drive forward the development and validation of rapid methods for X-ray structure determination, protein model building, refinement and structure validation. X-ray crystallography has become a central tool in modern drug and target discovery, providing important insights into molecular interactions and biological function. The past few years have seen many advances in the methods underlying macromolecular crystallography such as protein production, crystallization, cryo-crystallography and synchrotron technology. Together, these advances mean that X-ray data can be collected extremely quickly for many different crystals and ligand-bound complexes. The challenge is to ensure rapid and accurate interpretation of the data to provide valuable structural information.

*Corresponding Author

Recently, there has been tremendous interest in the use of direct methods for phase determination for macromolecules. This surge of interest has primarily resulted from two factors: the ability to obtain atomic resolution data in favorable cases and the development of powerful phasing methods including traditional direct methods so called half-baked and combinations of direct methods with isomorphous replacement and/or anomalous scattering [1]. Attempts have long been made to resolve the phase ambiguity arising from single-wavelength anomalous scattering (SAS) without using additional multi-wavelength or isomorphous derivative diffraction data. Multi-wavelength anomalous diffraction approach (MAD) generally requires a minimum of three wavelengths and the development of SAS is, therefore, highly significant given the explosion of synchrotron-based structural biology research. SAS experiment is straight forward and data can be collected in the standard way. There has recently been a great deal of interest using single-wavelength anomalous diffraction data in the elucidation of macromolecular structures [2,3], with investigations showing

that the SAS technique may be applied to many diverse problems, ranging from weak anomalous signals to highly complex substructures. Once experimental intensity data have been collected and processed, in the majority of cases, structure determination using the SAS technique proceeds *via* a three-step process. Firstly, the determination of the positions of the anomalous scatterers is carried out; phases are then developed in order to produce electron-density maps and in the final stage, these are interpreted using either manual or automatic methods to produce a starting model for refinement procedures [4,5].

2. Description of the program

Figure 1 shows the flow chart of the present work.

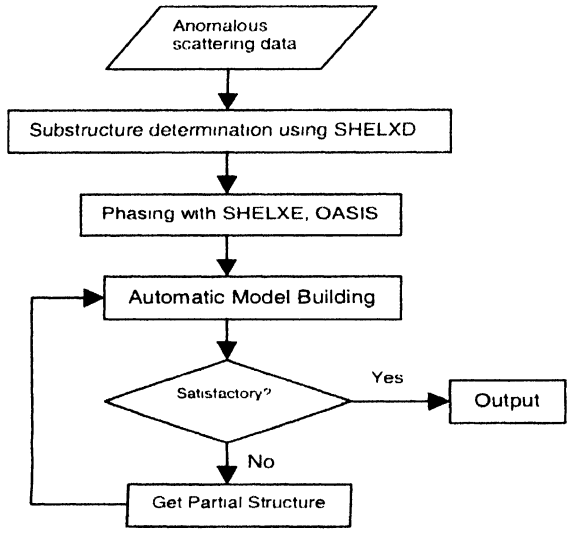

Figure 1. Flowchart

Information on anomalous scattering is important for the determination of protein structures. However, the single-wavelength anomalous-scattering method yields two possible solutions to each reflection which is known as the problem of phase ambiguity. If a method can be found to resolve the ambiguity, the SAS method would be useful technique in protein crystallography, since it is possible to solve a protein structure by either skipping the step of heavy-atom-derivative preparation if it contains suitable anomalous scatterers, or using only a heavy-atom derivative which may not be isomorphous to the native protein. Attempts that have been made to resolve the phase ambiguity arising from the SAS technique by direct methods since 1960's have succeeded in deriving a large number of three-phase structure invariants from the error-free data of a model protein structure [6].

The phase problem is reduced to a sign problem once the anomalous scatterers or the replacing heavy atom sites are located. OASIS is a computer program and it works on a direct-method procedure to break the phase ambiguity intrinsic to one-wavelength anomalous scattering (OAS) or single isomorphous replacement data [7]. All Friedel pairs (including centric reflections) were evaluated. It adopts the CCP4 format and has been written in Fortran 77. It is available in the CCP4 suite. The X-ray diffraction data and heavy atom site are the inputs for the program. The E-values are calculated based on the scale and temperature factors obtained from the Wilson plot and the absolute values of the phase doublets are calculated at this stage. Then for each reflection h, sigma2 relationships h and $\bar{h}$ are found and stored. The probability of phase doublets being positive and the best phase are calculated and then the figure of merit associated with every phase value is calculated. The resulting phase sets are further subjected into density modification.

The structure solution program SHELXD is useful for locating the heavy atoms or anomalous scatterers from SIR, SAD, SIRAS or MAD data. It is iterative dual-space direct methods based on phase refinement in reciprocal space and peak picking in real space. SHELXD locates relatively large numbers of anomalous scatterers efficiently from MAD or SAD data. Truncation of the data at a particular resolution in the range 3.0 - 3.5 Å can be critical to success. The efficiency can be improved by typically an order of magnitude by Patterson-based seeding instead of starting from random phases or sites [8].

The program SHELXE can read the heavy atom sites written by SHELXD and estimates the native phases and corresponding weights (figures of merit). SHELXE outputs the phases in an XtalView format. The map can be viewed using iterative graphics of the phases which can be improved by density modification. SHELXE was designed to provide a simple, fast and robust route from substructure sites found by the program SHELXD to an initial electron density map, if possible with an indication as to which heavy-atom enantiomorph is correct. The new sphere of influence algorithm combined with fuzzy solvent boundary enables some chemical knowledge to be incorporated into the density modification in a general and effective manner. In the special cases of high solvent content or very high-resolution data, high quality maps can be produced [9].

The phases obtained from SHELXE and OASIS are of superb quality to allow automated model building to be carried out using APR/wARP [10] followed by the refinement program REFMAC [11]. Attempts are here made in extending the applications to the high throughput structure elucidation with 1.7 Å resolution anomalous scattering data of thermolysin of approximately 34 kDa molecular weight and also for 2 Å truncated data obtained from it. In both cases, the starting is based on one zinc position obtained using SAS data. Application is also made with 1.45 Å resolution anomalous scattering data of glucose isomerase of approximately 44 kDa molecular weight using one manganese position obtained from the SAS data. These heavy

atom f ions were revealed by SHELXD. All the computations
ments here are carried out using the Pentium IV PC.

3. O view of the method

Anom ous scattering data from two known proteins,
therm sin and glucose isomerase, were used as test samples.

3.1. Fo molysin :

The di ction data were collected at a temperature of 100 K on the X9f synchrotron beamline at the National Synchrotron Light Source (Brookhaven National Laboratory, USA) using the ADSC Quantum4 CCD detector. This enzyme contains 316 residues, one Zn site and four calcium ions. Table 1a shows the crystallographic details of this protein for 1.7 Å data and 2.0 Å truncated data.

Table 1a. Details of the crystallographic data of thermolysin

For 1.7 Å data	
a (Å)	92.748
b (Å)	92.748
c (Å)	129.334
α (°)	90
β (°)	90
γ (°)	120
Space group	$P6_122$
Resolution range (Å)	20-1.7 (1.756-1.7)
Completeness (%)	98.5 (96.1)
$I/\sigma(I)$	53.5 (14.9)
Least value of anomalous signal-to-noise ratio	1.81
For 2.0 Å truncated data	
Resolution range (Å)	20-2.0 (2.066-2.0)
Completeness (%)	96.74 (93.94)
$I/\sigma(I)$	52.4 (29.89)
Least value of anomalous signal-to-noise ratio	2.29

The position of the anomalous scatterers in this enzyme (Zn) was located by direct methods program SHELXD. It gives three positions with a Correlation Coefficient (hereafter CC) value of 51.52. The top most peak was given to SHELXE for phasing and the CC has increased to 74.74. A map was calculated for the SHELXE output phases which showed 6043 peaks which were above the 3σ cut-off. The phases were then fed to ARP/wARP and REFMAC. After the initial model was refined, ten cycles of auto-building using ARP/wARP along with five cycles of REFMAC in each auto-building cycle were performed. Finally, ARP/wARP was able to build 310 out of 316 residues in three chains and has located 676 dummy atoms. At this stage, the R_{work} (hereafter, R_w) and R_{free} (hereafter, R_f) values were 16.0 and 21.0% respectively. The map also showed the densities in the missing region, so the manual model building was carried out for the missing residues. After the manual model building, 20 cycles of maximum-likelihood refinement were performed using REFMAC and solvent atoms were updated after the refinement using ARP/wARP 'build solvent atoms' script. The final R_w and R_f values were 17.7 and 20.4%, respectively. The average thermal factor (hereafter, B factor) for the current model is 14.9 Å^2. The backbone of this final model was superimposed with the reported model (PDB 1FJQ). The root-mean square deviation is 0.307 Å and all these details are shown in Table 1b. The Map Correlation Coefficient (MCC) between the SHELXE map and final map is 0.7704. Figure 1a shows the final cartoon diagram of this enzyme. Figure 1b shows a section of the final model superposed with the electron density of SHELXE map and also the final $2|F_o| - |F_c|$ map.

Figure 1a. Input E SHELXD peak to SHELXE Auto Built 310 residues

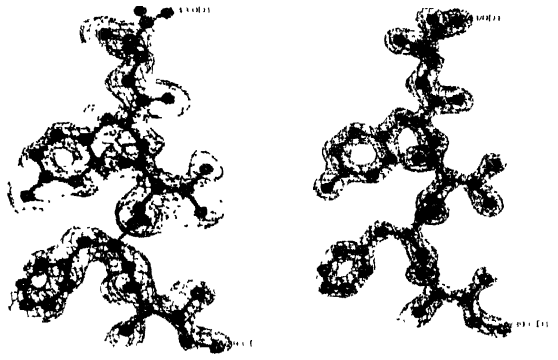

Figure 1b. Final model superposed with SHELXE map and final $2|F_o| - |F_c|$ map at 1σ

Truncated data of 2 Å resolution of this enzyme was prepared from SCALEPACK2MTZ option in CCP4 using 1.7 Å data and SHELXD gave three positions with a CC value of 54.20. The top most peak was given to SHELXE for phasing and the CC has increased to 69.45. The SHELXE map showed 4003 peaks which were greater than 3σ cut-off. The phases were then fed to ARP/wARP and REFMAC. Ten cycles of auto-building along with five cycles of REFMAC in each auto-building cycle were performed. Finally, ARP/wARP was able to build 311 out of 316 residues in four chains and has located 618 dummy atoms. At this stage, the R_w and R_f values were 14.5 and 22.5%, respectively.

Table 1d. Details of SHELXD, OASIS, ARP/wARP and REFMAC results for 1.7 Å data of thermolysin

OASIS input				
ATOM1 Zn	0.88054 0.54905 0.05460 1.00 0.0 BFAC 20.000 (output from SHELXD)			
PROGRAM	Resolution limit		20-1.7	
SHELXD	Three peaks	CC = 51.52		CC(weak) = 32.67
OASIS	One peak	472.8 Sec	2090 Peaks	MCC =
	Initial	R_w= 45.2		R_f=46.3
	No. of auto building cycles		10	
	No. of Refmac cycles in each auto building cycle		5	
ARP/wARP – I	**Final**	R_w= 28.7		R_f= 47.9
	Connectivity index		0.76	
	No. chains		9	
	No. res. built		75	
	No. of dummy atoms		2582	
	Initial	R_w= 28.7		R_f =47.9
	No. of auto building cycles		10	
	No. of Refmac cycles in each auto building cycle		5	
ARP/wARP – II	**Final**	R_w= 31.2		R_f= 47.4
	Connectivity index		0.83	
	No. chains		17	
	No. res. built		185	
	No. of dummy atoms		1811	
	Initial	R_w = 31.2		R_f =47.3
	No. of auto building cycles		10	
	No. of Refmac cycles in each auto building cycle		5	
ARP/wARP- III	**Final**	R_w= 16.9		R_f= 21.8
	Connectivity index		0.98	
	No. chains		3	
	No. res. built		309	
	No. of dummy atoms		705	
	R_{work} and R_{free} without dummy atoms	R_w = 27.5		R_f = 28.6
Final model with solvent atoms [solvent building carried out using 20 cycles of ARP/wARP building solvent atoms script]		R = 18.1		R_f = 20.5
		r.m.s. deviation of backbone atoms(1FJQ) 0.340 Å		

occupied by Mn^{2+} ion and the other by Mg^{2+}. The data was collected at a wavelength of 0.98 Å and belongs to I222 space group. The K X-ray absorption edge of manganese lies at 1.90 Å and at the wavelength used in this experiment, the imaginary component of the anomalous scattering (f'') of manganese varies between 2.8 and 1.3 electron units. The strongest anomalous scattering is provided by Mn, especially at shorter wavelengths where the anomalous effect of sulfur is very small. Table 2a shows the crystallographic details of this protein.

The location of the anomalous scatterers in this enzyme (Mn^{2+}) was performed by direct methods program SHELXD. SHELXD gave three positions with a *CC* value of 29.69. The top most peak was given to SHELXE for phasing and the final *CC* value was 81.79. A map was calculated for the SHELXE output phases and we were able to find 1812 peaks which were above the 3σ cut-off. The phases were then fed to ARP/wARP and REFMAC. Ten cycles of auto-building along with five cycles of REFMAC in each auto-building cycle were performed. Finally ARP/wARP was able to build 384 out of 388 residues in two chains and has located 902 dummy atoms. At this stage, the R and R_f values were 16.8 and 20.5%, respectively. The map also

Table 2a. Details of the crystallographic data of glucose isomerase

a (Å)	92.812
b (Å)	97.684
c (Å)	102.682
$\alpha = \beta = \gamma$(°)	90
Space group	I222
Resolution range (Å)	20-1.45 (1.51-1.45)
Completeness (%)	100.0 (100.0)
Rmerge	4.7 (11.)
I/σ (I)	62.3 (29.)
Least value of anomalous signal-to-noise ratio	1.22

showe ie densities in the missing region and manual model buildi was carried out for the missing residues. After the manua model building, 20 cycles of maximum-likelihood refine it were performed using REFMAC and solvent atoms were a ated after the refinement using ARP/wARP 'build solven oms' script. The final R_w and R_f values were 17.3 and 18.9% spectively. The average B factor for the current model is 9.7 The backbone of this final model was superimposed with th ne in $P2_12_12$ space group of this enzyme (PDB 1OAD). The roo mean square deviation is 0.184 Å and all these details are shown in Table 2b. The Map Correlation Coefficient (MCC) between the SHELXE map and final map is 0.8127. Figure 2a shows the final cartoon diagram of this enzyme. Figure 2b shows a section of the final model superposed with the electron density of SHELXE map and also the final $2|F_o| - |F_c|$ map.

As an alternative method, OASIS was run for the top most peak obtained from SHELXD. Density modification (DM) using the CCP4 program was then carried out with the resulting phase sets. A map was calculated for the OASIS output phases and we were able to find 908 peaks which were above the 3σ cut-off. The automated model building was performed using ARP/

Table 2b. Details of SHELXD, SHELXE, ARP/wARP and REFMAC results for glucose isomerase

SHELXD output

Mn01 1 0.583054 0.133270 0.066371 1.0000 0.2
Mn02 1 0.631714 0.147301 0.080120 0.2927 0.2
Mn03 1 0.612625 0.175293 0.241702 0.2350 0.2

SHELXE input

Mn01 1 0.583054 0.133270 0.066371 1.0000 0.2

PROGRAM	Resolution limit		20-1.45	
SHELXD	3 peaks	CC = 29.69	CC(weak) = 19.16	
SHELXE	1 peak	CC = 81.79	1812 peaks	MCC = 0.8127
	Initial	**R_w = 33.5**		**R_f = 32.7**
	No. of auto building cycles		10	
	No. of Refmac cycles in each auto building cycle		5	
ARP/wARP	**Final**	R_w = 16.8		R_f = 20.5
	Connectivity index		0.99	
	No. chains			
	No. res. built		384	
	No. of dummy atoms		902	
	R_{work} and R_{free} without dummy atoms	**R_w = 26.1**		**R_f = 26.9**
Final model with solvent atoms [solvent building carried out using 20 cycles of ARP/wARP. building solvent atoms script]		**R_w = 17.3**		**R_f = 18.9**
		r.m.s. deviation of backbone atoms(1OAD) 0.184 Å		

Table 2c. Details of SHELXD, OASIS, ARP/wARP and REFMAC results for glucose isomerase.

OASIS input

ATOM1 Mn 0.58305 0.13327 0.06637 1.00 0.0 BFAC 20.000 (output from SHELXD)

PROGRAM	Resolution limit		20-1.45	
OASIS	One peak (from SHELXD)	1463.0 Sec	908 peaks	MCC = 0.2978
	Initial	**R_w= 47.7**	**R_f=47.8**	
	No. of auto building cycles		10	
	No. of Refmac cycles in each auto building cycle		5	
ARP/wARP	**Final**	R_w = 16.9	R_f = 20.3	
	Connectivity index		0.99	
	No. chains		2	
	No. res. built		385	
	No. of dummy atoms		878	
	R_{work} and R_{free} without dummy atoms	**R_w = 26.1**	**R_f = 27.0**	
Final model with solvent atoms [solvent building carried out using 20 cycles of ARP/wARP: building solvent atoms script]		R_w = 17.5	R_f = 19.3	
		r.m.s. deviation of backbone atoms(1OAD). 0.170 Å		

wARP for these modified phases. Finally, ARP/wARP was able to build 385 out of 388 residues in two chains with a connectivity

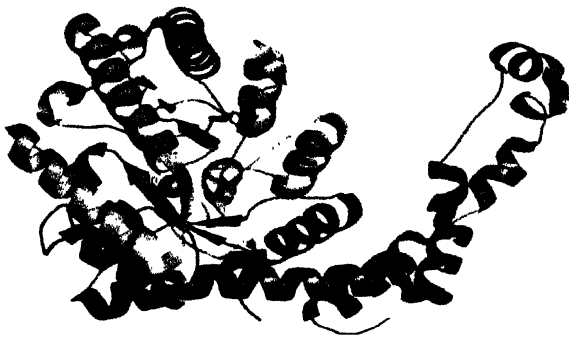

Figure 2a. Input 1 SHELXD peak to SHELXE Auto Built 384 residues

Figure 2b. Final model superposed with SHELXL map and final $2|F_o| - |F_c|$ map at 1σ.

index of 0.99. At this stage, the R_w and R_f values were 16.9 and 20.3%, respectively. Manual model building was carried out for the missing residues and solvent atoms were updated after the refinement using ARP/wARP 'build solvent atoms' script. The final R_w and R_f values were 17.5 and 19.3%, respectively. The average B factor for the current model is 9.5 Å^2. The backbone of this final model was superimposed with that of $P2_12_12$ form of this enzyme (PDB 1OAD). The root-mean square deviation is 0.170 Å and all these details are shown in Table 2c. The Map Correlation Coefficient (MCC) between the OASIS map and final map is 0.2978. Figure 2c shows the final cartoon diagram of this enzyme. Figure 2d shows a section of the final model superposed with the electron density of OASIS map and also the final $2|F_o| - |F_c|$ map.

4. Conclusion

The above work emphasizes the applicability of the SAS technique to solve a macromolecular structure when data extends to 2.0 Å resolution. Only one anomalous scatterer is used here. Many proteins host light metals such as calcium, manganese, potassium as cofactors or recruit them as stabilizing agents. These metals may provide an opportunity to bypass the preparation of heavy-atom derivatives or the incorporation of selenomethionine residues into native sequences and allow *de novo* crystal structure determination.

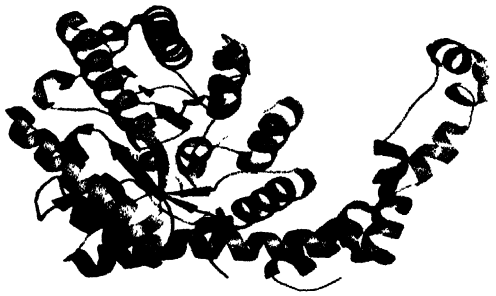

Figure 2c. Input 1 SHELXD peak to OASIS Auto Built 385 residues

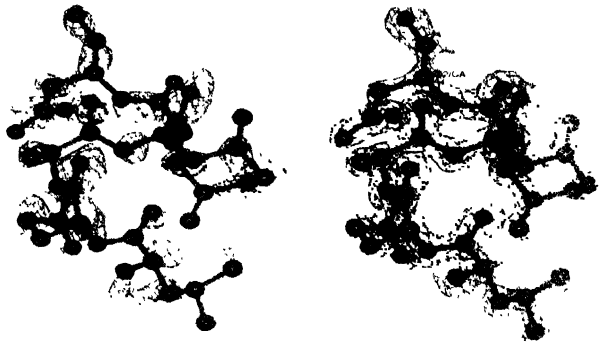

Figure 2d. Final model superposed with OASIS map and final $2|F_o| - |F_c|$ map at 1σ

The above results demonstrate that the direct method is capable of discriminating the correct phase in a bimodal distribution of a protein reflection by exploiting single-wavelength anomalous scattering diffraction data which extends to modest resolution. The combination of SAS data and direct methods is a powerful approach for resolving phases for protein structure determination; its wider adoption would result in a major saving of synchrotron-radiation experimental time, about 2/3rd. This work also adds substantial evidence that even with single-wavelength anomalous scattering data, a macromolecular structure can be solved with the existing sophisticated programs with the knowledge of just one anomalous scatterer and it is also seen from our above studies that the SHELXE phases are much better than OASIS phases, which is confirmed by map correlation coefficient, electron density maps and their output peaks. The SAS method could therefore, play an important role in the high-throughput complete automatic procedures currently planned for structural genomics initiatives.

Acknowledgments

SS thanks Council of Scientific and Industrial Research (CSIR) for providing Senior Research Fellowship. DV acknowledges Bioinformatics division of Department of Biotechnology (DBT) and University Grants Commission (UGC), Govt. of India for major projects supporting this work and thanks Venture Business Laboratory authorities, Nagoya Univeristy, Nagoya, Japan for

he VIS g Professorship assignments in short terms and also
ackno lges financial support to the Department under UGC-
SAP at)ST-FIST programmes. DV thanks Prof. Z. Dauter for
provid the anomalous scattering data sets of thermolysin
and gl se isomerase.

Refere ces

[1] Hauptman *Curr Opin Struct Biol* **7** 672 (1997)
[2] M Rice, T N Eamest and A T Brunger *Acta Cryst* **D56** 14[illegible]3 (2000)
[3] Dauter, M Dauter and E J Dodson *Acta Cryst.* **D58** 494 (2002)
[4] Gdupi A Ramagopal, M Dauter and Z Dauter *Acta Cryst* **D59** 8[illegible]8 (2003)
[5] Gordon A Leonard, G Sainz, Maarike, M E de Backer and S McSweeney *Acta Cryst* **D61** 388 (2005)
[6] F Hai-Fu, H Quan, G Yuan-xin, Q Jin-zi and Z Chao-de *Acta Cryst.* **A46** 935 (1990)
[7] Q Hao, Y X Gu, C D Zheng and H F Fan *J Appl Cryst.* **33** 980 (2000)
[8] Thomas R Schneider and George M Sheldrick *Acta Cryst* **D58** 1772 (2002)
[9] G M Sheldrick *Z Kristallogr* **217** 644 (2002)
[10] A Perrakis, R Morris and V S Lamzin *Nature Struct. Biol* **6** 458 (1999)
[11] G N Murshudov, A Lebedev, A A Vagin, K S Wilson and E J Dodson *Acta Cryst* **D55** 247 (1999)